M2 SES-IES Économétrie des variables qualitatives

Examen du 12 mai 2015

Exercice 1

Cet exercice porte sur les données observées sur un échantillon de 474 employés tirés au sort dans une entreprise canadienne. Les variables étudiées ici sont les suivantes :

- `salary` (salaire brut actuel en $ par an)
- `salbegin` (salaire de départ en $ par an)
- `jobtime`(nombre de mois depuis l'entrée dans l'entreprise)
- `prevexp` (nombre de mois de travail avant l'entrée dans l'entreprise)
- `educ` (nombre d'années d'étude)
- `sex` (sexe à deux modalités H = Homme et F = Femme)

On souhaite expliquer la variable `salary` en fonction de toutes les autres variables (`salbegin`, `jobtime`, `prevexp`, `educ` et `sex`) à l'aide de la régression linéaire.

1. Nous avons déterminé la matrice de corrélation.

```
              salary    salbegin      jobtime      prevexp        educ
salary    1.00000000  0.88011747  0.084092267 -0.097466926  0.66055891
salbegin  0.88011747  1.00000000 -0.019753475  0.045135627  0.63319565
jobtime   0.08409227 -0.01975347  1.000000000  0.002978134  0.04737878
prevexp  -0.09746693  0.04513563  0.002978134  1.000000000 -0.25235252
educ      0.66055891  0.63319565  0.047378777 -0.252352521  1.00000000
```

   (a) Indiquer pour quels couples de variables la corrélation linéaire observée est la plus forte, la plus faible. Que peut-on dire de la corrélation linéaire entre le salaire de départ et le salaire actuel ?

   (b) Pourquoi n'y a-t-il pas la variable `sex` dans le matrice de corrélation ?

2. Nous avons ajusté un modèle de régression linéaire multiple expliquant `salary` en fonction de toutes les autres (`salbegin`, `jobtime`, `prevexp`, `educ` et `sex`).

   Modèle 1 :

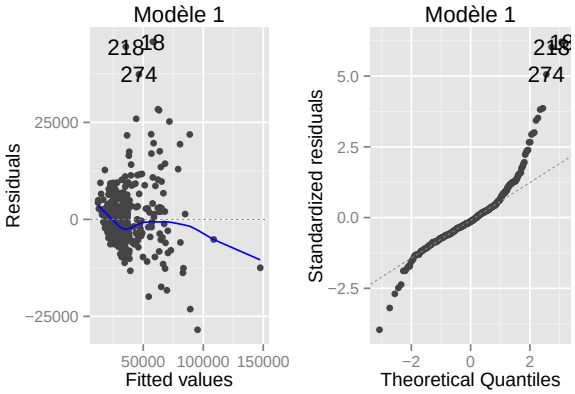
```
Modele1=lm(formula = salary ~ salbegin + jobtime + prevexp + educ + sex,
data = Salaire)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.255e+04  3.475e+03  -3.612 0.000337 ***
salbegin     1.723e+00  6.051e-02  28.472  < 2e-16 ***
jobtime      1.545e+02  3.408e+01   4.534 7.37e-06 ***
prevexp     -1.944e+01  3.583e+00  -5.424 9.36e-08 ***
educ         5.930e+02  1.666e+02   3.559 0.000410 ***
sexF        -2.233e+03  7.921e+02  -2.819 0.005021 **
---
Residual standard error: 7410 on 468 degrees of freedom
Multiple R-squared:  0.8137,Adjusted R-squared:  0.8117
F-statistic: 408.7 on 5 and 468 DF,  p-value: < 2.2e-16
```

   (a) Tester la significativité globale du modèle à un niveau de risque de 5% en n'oubliant pas de donner les hypothèses nulles et alternatives du test. Que peut-on conclure ?

   (b) Relever et interpréter la valeur observée du coefficient R^2.

(c) Quelles sont les variables significatives au seuil de signification de 5% ?

(d) Que représentent les graphes ci-dessous ?



(e) Pensez vous que le modèle ajusté est pertinent ? Justifier.

3. Nous avons appliqué une transformation logarithmique aux variables `salary` et `salbegin` et nous avons ajusté un modèle de régression linéaire multiple en remplaçant ces variables par les variables transformées.

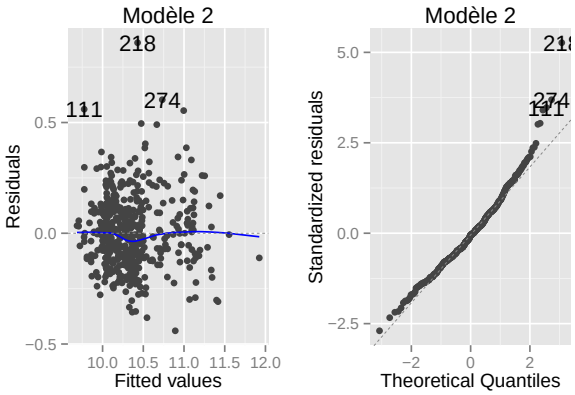
Modèle 2 :

```
Modele2=lm(formula = log(salary) ~ log(salbegin) + jobtime + prevexp +
    educ + sex, data = Salaire)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.116e+00  3.125e-01   3.571 0.000392 ***
log(salbegin)  9.107e-01  3.382e-02  26.924  < 2e-16 ***
jobtime        4.517e-03  7.579e-04   5.960 4.97e-09 ***
prevexp       -5.527e-04  7.932e-05  -6.968 1.10e-11 ***
educ           1.071e-02  3.912e-03   2.737 0.006431 **
sexF          -4.995e-02  1.844e-02  -2.708 0.007019 **
---
Residual standard error: 0.1639 on 468 degrees of freedom
Multiple R-squared:  0.8317,Adjusted R-squared:  0.8299
F-statistic: 462.6 on 5 and 468 DF,  p-value: < 2.2e-16
```

(a) Tester la significativité globale du modèle à un niveau de risque de 5% en n'oubliant pas de donner les hypothèses nulles et alternatives du test. Que peut-on conclure ?

(b) Relever et interpréter la valeur observée du coefficient R^2.

(c) Peut-on valider ce modèle ? Expliquer.



4. Lequel des deux modèles de régression multiple considérés préférez vous ? Appuyez votre réponse sur les graphes pertinents.
5. Donner l'équation du modèle ajusté que vous avez choisi et préciser l'interprétation des coefficients estimés.

Exercice 2

Rappel : dans le modèle logistique, nous cherchons à expliquer une variable Y, qui vaut 0 ou 1, à partir d'une variable explicative X (ou d'un vecteur de variables explicatives également noté X). Nous modélisons pour cela $\pi(x) = \mathbb{P}(Y = 1|X = x)$ par

$$\pi(x) = \frac{\exp(\beta_0 + \langle \beta, x \rangle)}{1 + \exp(\beta_0 + \langle \beta, x \rangle)}$$

où $\langle \beta, x \rangle = \beta_1 x$ si X est une variable simple et $\langle \beta, x \rangle = \beta_1 x_1 + \beta_2 x_2$ si X est un vecteur à deux composantes. Ce modèle est équivalent à $\log(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \langle \beta, x \rangle$ avec β_0 et β inconnus. Enfin, dans le cas où X prend également les valeur 0 ou 1, on montre que $\hat{\beta}_0 = \log(n_{21}/n_{11})$ et $\hat{\beta}_1 = \log((n_{11}n_{22})/(n_{12}n_{21}))$ où n_{ij} représente l'effectif observé pour $i = 1, 2$ et $j = 1, 2$ du tableau de contingence correspondant :

	X=0	X=1
Y=0	n_{11}	n_{12}
Y=1	n_{21}	n_{22}

Nous traitons un problème de défaut bancaire. La variable `default` est la variable à expliquer. Nous disposons ici d'un échantillon de taille 10000 et deux variables explicatives `student` et `balance`.

— `default` : `Yes` (ou 1) si le client fait défaut sur sa dette et `No` (ou 0) sinon.
— `student` : `Yes` (ou 1) si le client est un étudiant et `No` (ou 0) sinon
— `balance` : montant moyen mensuel d'utilisation de la carte de crédit

1. On considère pour commencer un modèle de régression logistique simple où on cherche à expliquer `default` en fonction de `student`.

   (a) À l'aide du tableau de contingence ci dessous, calculer "à la main" les coefficients estimés du modèle logistique.

   ```
          student   No  Yes
   default
   No               6850 2817
   Yes               206  127
   ```

   (b) Donner l'équation du modèle logistique ajusté.

   (c) Calculer le rapport de rapport de chances (odds ratio). Que peut-on en conclure ?

2. Nous avons utilisé le logiciel R pour ajuster ce même modèle logistique simple.

   Modèle 1 :

   ```
   Modele1 = glm(formula = default~student, family = binomial(link = "logit"),
   data = Default)
   Coefficients:
               Estimate Std. Error z value Pr(>|z|)
   (Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
   studentYes   0.40489    0.11502    3.52 0.000431 ***
   ---
   AIC: 1600.5
   ```

(a) Donner les coefficients estimés par le logiciel R du modèle logistique ajusté.

(b) Comparer avec les coefficients estimés dans la question précédente.

3. Nous avons ajusté un modèle logistique multiple où on cherche à expliquer `default` en fonction de `student` et de `balance`.

Modèle 2 :

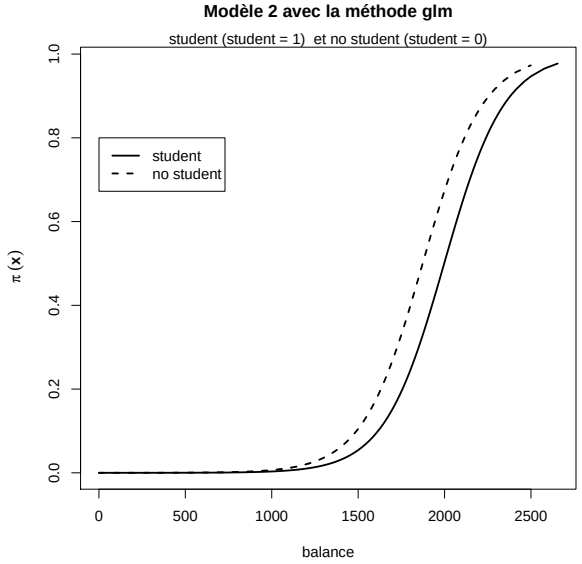
```
Modele2=glm(formula = default ~ student + balance,
family = binomial(link = "logit"), data = Default)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.749496  3.692e-01 -29.116  < 2e-16 ***
studentYes  -0.7148776  1.475e-01  -4.846 1.26e-06 ***
balance      0.0057381  2.318e-04  24.750  < 2e-16 ***
---
AIC: 1577.7
```

(a) Donner l'équation du modèle logistique ajusté avec les coefficients estimés pour les `"student=Yes"` et pour les `"student=No"`.

(b) Nous avons relevé les valeurs estimées de la proportion de `default` selon les caractéristiques de trois clients au hasard. Est-ce qu'on peut dire si les clients 1, 137 et 9999 feront `default` ? Justifier.

```
predict(Modele2,newdata=Default[c(1,137,9999),c("student","balance")],
type="response")
          1         137        9999
0.001409096 0.050602655 0.148507089
```

(c) Commenter **brièvement** la figure ci-dessous.



4. Utiliser le critère `AIC` pour choisir un modèle. Lequel choisissez-vous ? Justifier